

Latent semantics of action verbs reflect phonetic parameters of intensity and emotional content

Michael Kai Petersen*

Cognitive Systems
DTU Compute, Building 324
Technical University of Denmark
DK-2800 Kgs.Lyngby, Denmark
* mkai@dtu.dk

Abstract

Conjuring up our thoughts, language reflects statistical patterns of word co-occurrences which in turn come to describe how we perceive the world. Whether counting how frequently nouns and verbs combine in Google search queries, or extracting eigenvectors from term document matrices made up of Wikipedia lines and Shakespeare plots, the resulting latent semantics capture not only the associative links which form concepts, but also spatial dimensions embedded within the surface structure of language. As both the shape and movements of objects have been found to be associated with phonetic contrasts already in toddlers, this study explores whether articulatory and acoustic parameters may likewise differentiate the latent semantics of action verbs. Selecting 3×20 emotion, face, and hand related verbs known to activate premotor areas in the brain, their mutual cosine similarities were computed using latent semantic analysis LSA, and the resulting adjacency matrices were compared based on two different large scale text corpora; HAWIK and TASA. Applying hierarchical clustering to identify common structures across the two text corpora, the verbs largely divide into combined mouth and hand movements versus emotional expressions. Transforming the verbs into their constituent phonemes, and projecting them into an articulatory space framed by tongue height and formant frequencies, the clustered small and large size movements appear differentiated by front versus back vowels corresponding to increasing levels of arousal. Whereas the clustered emotional verbs seem characterized by sequences of close versus open jaw produced phonemes, generating up- or downwards shifts in formant frequencies that may influence their perceived valence. Suggesting, that the latent semantics of action verbs reflect parameters of intensity and emotional polarity that appear correlated with the articulatory contrasts and acoustic characteristics of phonemes.

Introduction

If language adapted to the brain like a virus [1], constrained by sensorimotor circuits linking articulatory gestures with aspects of motion [2], then parameters of size and intensity might also potentially be reflected in the latent semantics of words. Spatiotemporal metaphors are ubiquitous in phrases like ‘hitting the road’, ‘christmas is approaching’ or ‘thinking out of the box’, where we reinterpret ourselves as objects that are subject to forces of gravity or moving along virtual time lines [3]. Aspects of motion appear to constrain how we internally represent affordances for potential action, as perceptual states are reenacted from memory traces formed by sensorimotor circuits [4]. Likewise, emotions are metaphorically constrained by spatial parameters; we process positive and negative words faster depending on whether they are presented at the top or bottom of a screen, as upwards or downwards is perceived as vertically congruent with pleasant and unpleasant connotations respectively [5]. Adding to a growing amount of evidence for embodied cognition [6], where not only action verbs like ‘push’ are associated

with trajectories, but also terms like ‘argue’ and ‘respect’ appear to be grounded in a conceptual space framed by horizontal and vertical axes [7]. Recent studies have demonstrated that spatial dimensions can be retrieved from the surface structure of sentences describing horizontal and vertical movements [8]. Similarly that it is feasible to geographically map out the relative distances between cities based on how they as words co-occur in news articles [9] or in fiction like “Lord of the Rings” [10]. Underlying parameters of size and intensity may even be reflected in the phonetic building blocks of language, as behavioral studies have shown that high front vowels are perceived as lighter and associated with smaller organisms than words involving back vowels [11]. Correspondences between articulatory gestures and the shapes of objects have been found already in toddlers, who associate back produced vowels as in ‘bouba’ with rounded forms whereas bright front vowels as in ‘kiki’ are associated with edgy outlines [12]. Phonological cues might thus provide a semantic bootstrapping that would facilitate language learning [13]. If aspects of action based language have through Hebbian learning been associated with sequences of verbs and nouns [2], the underlying parameters of motion could potentially also be retrieved from the latent semantics of action verbs. To explore whether spatiotemporal parameters might also be reflected in the phonetic structure of action verbs, 3 × 20 emotion, face, and hand related verbs were selected, which have in earlier neuroimaging studies been shown to activate motor, premotor and prefrontal circuits in the brain during a passive reading task [14]. Applying latent semantic analysis LSA [15] to define the cosine similarities between the action verbs, adjacency matrices were computed based on two different large scale text corpora; HAWIK, consisting of 22829 words found in 67380 excerpts of Harvard Classics literature, Wikipedia articles and Reuters news [16] and TASA, consisting of 92409 words extracted from 37651 text excerpts, reflecting the educational material US students have been exposed to when entering their first year of college [17]. To identify significant structures among the action verbs, hierarchical clustering was applied to the two adjacency matrices derived from the HAWIK and TASA corpora, using multistep-multiscale bootstrap resampling with Pearson correlation as distance measure [18]. Selecting the action verbs which were grouped similarly based on both the HAWIK and TASA text corpora, the words were annotated with their corresponding user rated word norms related to the psychological dimensions of valence and arousal [19], thus defining their emotional polarity and perceived intensity [20]. Subsequently the action verbs were transformed into ARPAbet phonemes using the CMU text to speech pronunciation dictionary [21], and their primary stress vowels projected into an articulatory space defined by tongue height and front-back position in the international phonetic alphabet (IPA) and framed by their corresponding F1 and F2 formant frequencies [22].

Results and Discussion

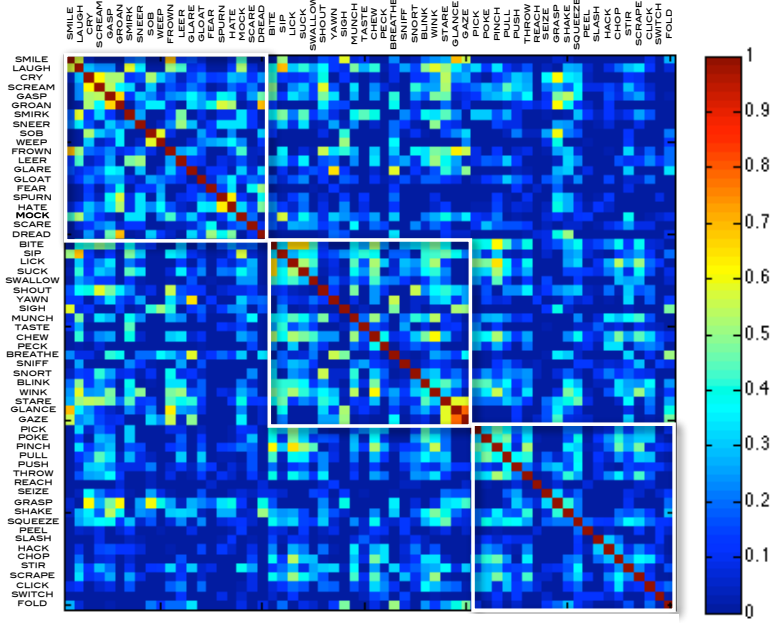


Figure 1. HAWIK adjacency matrix of emotion, face and hand action verbs, defining their cosine similarities generated by applying LSA latent semantic analysis and SVD singular value decomposition to reduce the dimensionality to the 125 most significant eigenvalues. The latent semantic relations between the 3×20 verbs are computed based on the HAWIK text corpus term document co-occurrence matrix, consisting of 22829 words extracted from 67380 excerpts of Harvard Classics literature, Wikipedia articles and Reuters news.

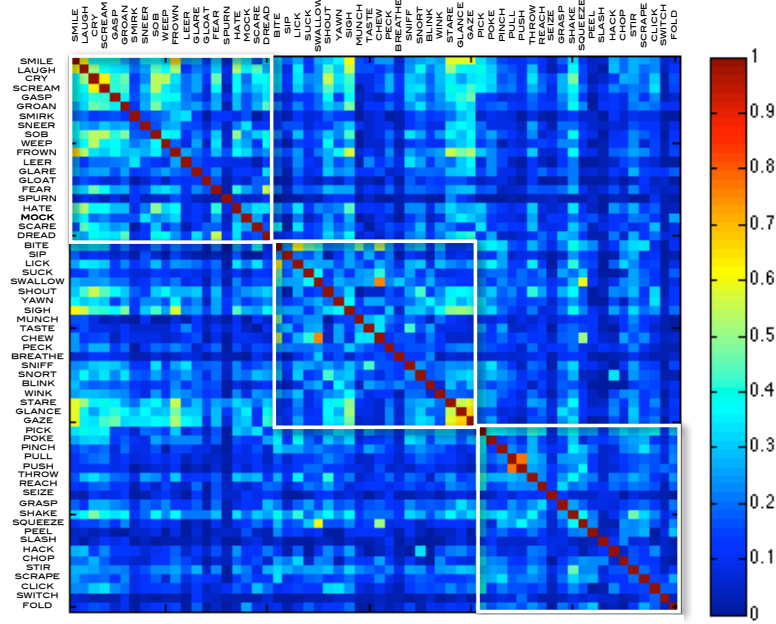


Figure 2. TASA adjacency matrix of emotion, face and hand action verbs, defining their cosine similarities generated by applying LSA latent semantic analysis and SVD singular value decomposition to reduce the dimensionality to the 300 most significant eigenvalues. The latent semantic relations between the 3×20 verbs are computed based on the TASA text corpus term document co-occurrence matrix, consisting of 92409 words extracted from 37651 text excerpts, reflecting the educational material US students have been exposed to when entering their first year of college.

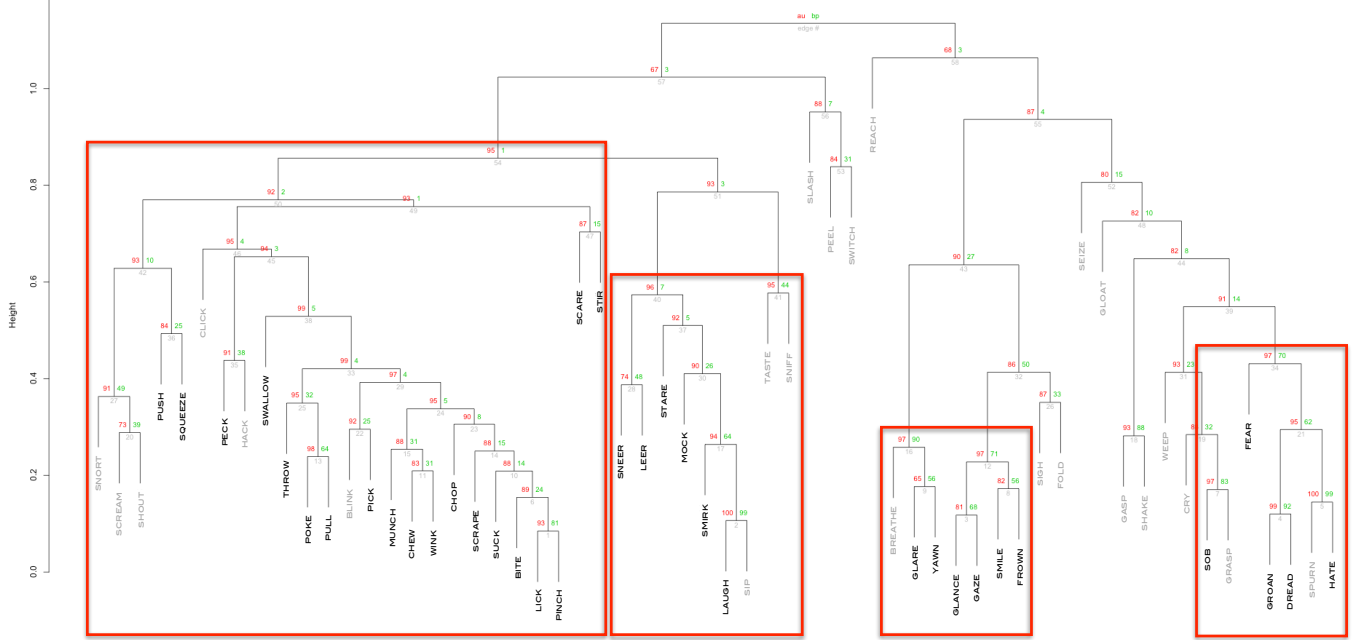


Figure 3. Hierarchical clustering of action verbs based on HAWIK adjacency matrix, using multistep-multiscale bootstrap resampling ($n = 100,000$) with Pearson correlation as distance measure. AU approximately unbiased p – values ($p \leq 0.05$, $SD \leq 0.004$) define significant clusters as red rectangles (bold fonts denoting verbs similarly clustered based on the TASA dendrogram) from left to right: 1. Mouth and hand movements of low to high arousal (2.81 - 7.10, $M = 4.44$) 2. Emotional contrasting low versus high valence (3.30 - 7.53, $M = 4.45$) 3. Facial expressions of low arousal (2.63 - 4.52, $M = 3.83$) 4. Negative emotions of low valence (1.96 - 3.90, $M = 2.93$)

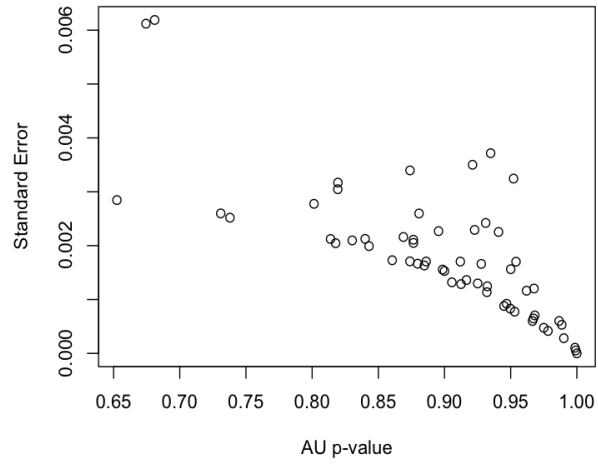


Figure 4. AU p – values plotted against standard errors for the clusters identified in the HAWIK adjacency matrix, computed using multistep-multiscale bootstrap resampling ($n = 100.000$), indicating standard errors ≤ 0.004 for the AU approximately unbiased p-values ≤ 0.05 used to reject the null hypothesis of the hierarchical clustering structure not being supported by the data.

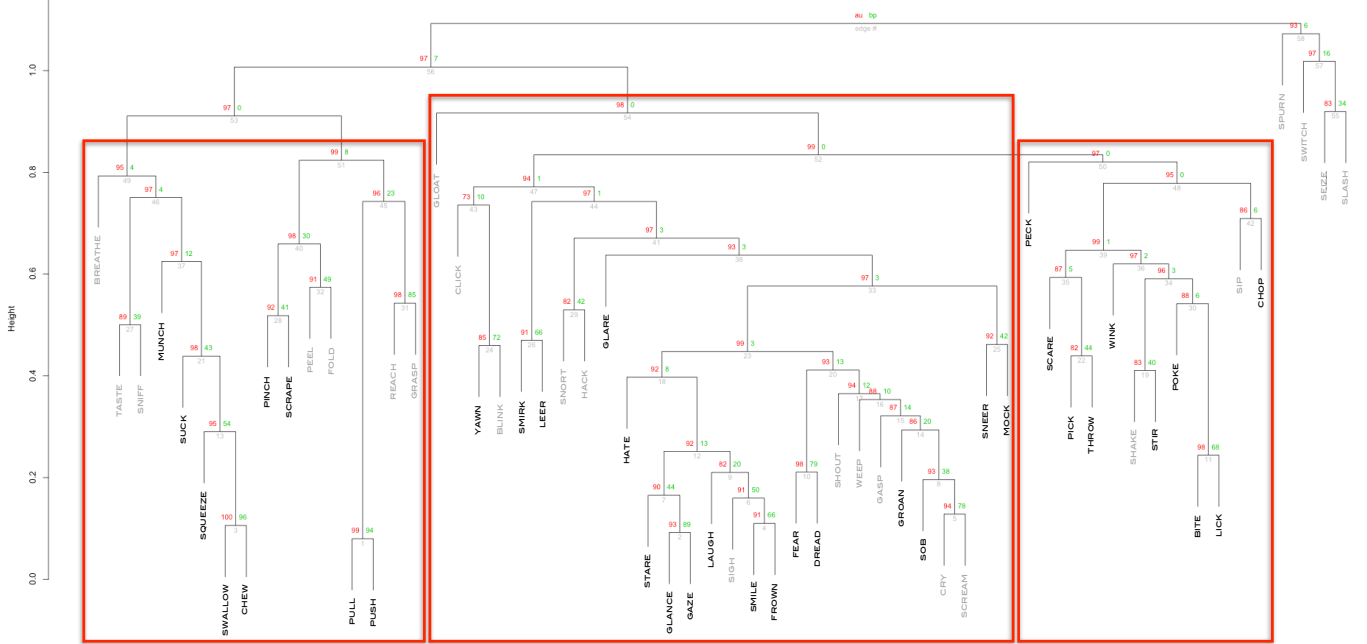


Figure 5. Hierarchical clustering of action verbs based on TASA adjacency matrix, using multistep-multiscale bootstrap resampling ($n = 100,000$) based on Pearson correlation with approximately unbiased AU p – values ($p \leq 0.05$, $SD \leq 0.007$) define significant clusters as red rectangles (bold fonts denoting verbs similarly clustered based on the HAWIK dendrogram) from left to right: 1. Mouth and hand movements of low to medium arousal (2.60 - 5.60, $M = 3.98$) 2. Emotional expressions of low versus high valence (1.96 - 7.89, $M = 3.95$) 3. Mouth and hand motion characterized by enhanced dominance (4.58 - 6.74, $M = 6.12$)

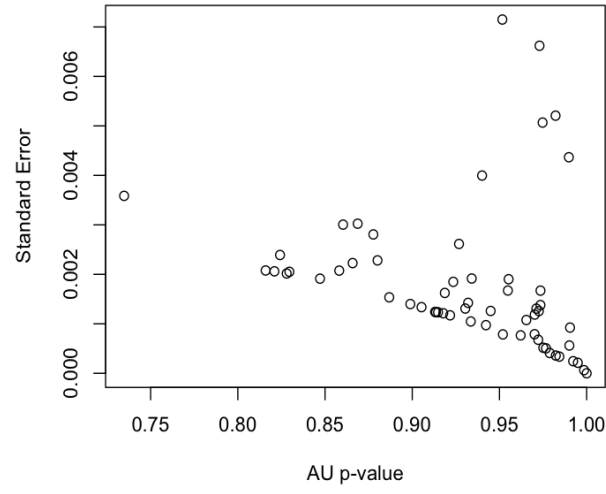


Figure 6. AU p – values plotted against standard errors for the clusters identified in the TASA adjacency matrix, computed using multistep-multiscale bootstrap resampling ($n = 100.000$), indicating standard errors ≤ 0.007 for the AU approximately unbiased p-values ≤ 0.05 used to reject the null hypothesis of the hierarchical clustering structure not being supported by the data.

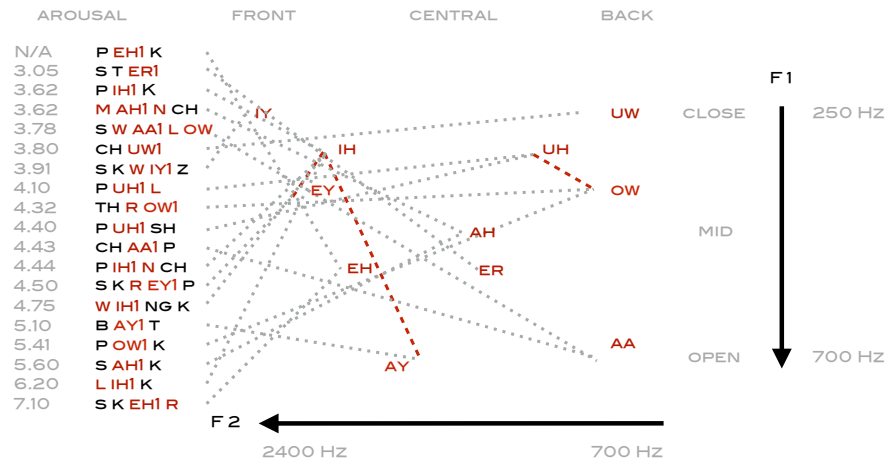


Figure 7. Articulatory projections of primary stress vowels in mouth and hand verbs horizontally differentiate front versus back vowels (dotted) and diphthongs (dashed) mapped according to tongue height, front-back position and rounding, whereas auditory features are defined by the corresponding F1 and F2 formant frequencies.

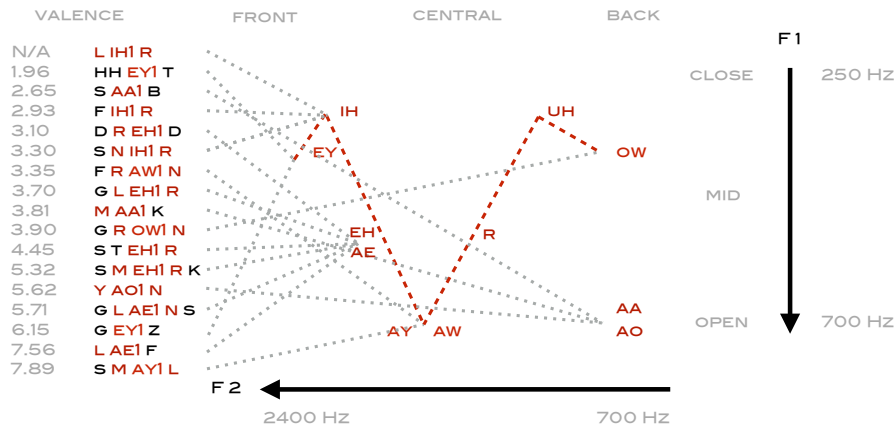


Figure 8. Articulatory projections of primary stress vowels in facial and emotional verbs of positive and negative valence, vertically contrast open versus close vowels (dotted) and diphthongs (dashed), mapped according to tongue height, front-back position and rounding, while auditory features are defined by the corresponding F1 and F2 formant frequencies.

Hierarchical structures

In the LSA adjacency matrices generated using the HAWIK (Fig. 1) and TASA (Fig. 2) text corpora, the co-occurrences of verbs in the upper left quadrant stand out. Here emotions such as ‘smile’ and ‘frown’ are coupled with facial expressions like ‘gaze’ and ‘glance’ further down along the diagonal, but only sparsely trigger hand related verbs such as ‘pick’ or ‘poke’ in the lower right corner of the matrix. Whereas verbs related to cyclical jaw and tongue motion such as ‘bite’ or ‘chew’ strongly co-activate hand movements like ‘pinch’ in the HAWIK or ‘squeeze’ in the TASA corpora. Behaviorally, combinations of hand and mouth movements typically co-occur in sequences when we pick up something using the fingers, rotating the elbow and moving the hand towards the face while simultaneously opening the mouth [23]. Such sequences of frequently co-occurring movements appear to form a vocabulary of hierarchically structured motor schemas, which are recursively combined into object oriented patterns of motion. In effect reducing the complexity when manipulating objects from a large number of free parameters to only a few dimensions of orientation and size. These parameters might in turn be related to stored representations of sequences constrained by physical parameters of distance and gravity [24] [25].

To quantify the hierarchical structures among the action verbs, characterized by heavily interconnected subgraphs which are only sparsely linked to other modules [26], hierarchical clustering was applied to the HAWIK and TASA adjacency matrices. Assessing to what degree the resulting hierarchical tree structures are supported by the data, approximately unbiased AU p-values were computed by multistep multiscale resampling changing the sample size of the bootstrap replicates [18] to define statistically significant clusters based on the HAWIK (Fig 3 & 4: $p \leq 0.05$, $SD \leq 0.004$) and TASA (Fig 5 & 6: $p \leq 0.05$, $SD \leq 0.007$) dendograms. Visualized in a dendogram, the x axis represents the connectivity between the pairwise most similar leaf nodes and the y axis defines their distances. The resulting tree thus approximates the structural properties of the data by grouping the nodes at specific levels within a hierarchy [27]. When comparing the dendograms to identify common structures across the two text corpora, clusters of combined mouth and hand movements are in both the HAWIK and TASA separated from the clusters of facial expressions and emotions. As exemplified by the action verbs in HAWIK cluster 1 describing small size finger precision grip combined with oscillatory jaw and tongue motion such as ‘squeeze’, ‘pinch’, ‘scrape’, ‘suck’, ‘chew’, ‘munch’ and ‘swallow’, which are similarly grouped in cluster 1 of the TASA dendogram. More forceful movements in HAWIK cluster 1, involving the arm combined with a whole hand grip like ‘stir’ ‘pick’ ‘throw’ ‘chop’ ‘poke’, plus oscillatory motion as in ‘bite’, ‘lick’, ‘peck’ and ‘wink’, are in the TASA dendogram grouped separately in cluster 3. Indicating, that motor schemas grouping small size hand and mouth movements involved when eating are also reflected in the latent semantics of action verbs across both the TASA and HAWIK text corpora. Likewise, the hierarchical relations between finger precision motion and more forceful whole hand grips involving the arm, are reflected in the hierarchical structures of action verbs juxtaposing small size gestures against larger movements. Suggesting, that hierarchical linguistic structures reflecting aspects of size and intensity might lower the number of free parameters in action verbs. Similar to how ideal antonym pairs like ‘pull’ / ‘push’ or ‘smile’ / ‘frown’ conceptually reduce dimensionality by sharing multiple semantic features while crucially differing along a single parameter [28].

Motion and emotion

Exploring how the clustered action verbs are perceived along a reduced number of psychological dimensions, they were subsequently annotated with crowd sourced user ratings available online from the “Norms of valence, arousal and dominance for 13915 English lemmas” data set [19]. Similar to the smaller ANEW data set [29] the user ratings assess how pleasant, intense and controlled the verbs are described as being on a scale from 1 to 9, along the psychological dimensions of valence, arousal and dominance [20]. Resulting in a user rated semantic space framed by three parameters: valence defining how pleasant or unpleasant something is perceived on a bipolar axis (1.26-8.48, $M = 5.06$, $SD = 1.68$), arousal capturing

the level of intensity going from calm to excited (1.60-7.79, $M = 4.21$, $SD = 2.30$) and dominance (1.68-7.74, $M = 5.18$, $SD = 2.16$) describing the degree of feeling in control. Rather than considering the verbs as individual terms they are thus represented as emotional buoys in a semantic space, related to aspects of approach or avoidance behavior that are psychologically grounded in motivational survival mechanisms [30] [31]. Combining the user rated annotations on how the words are perceived with the clustering analysis, the hierarchical structures of mouth and hand motion verbs in the HAWIK cluster 1 appears to a larger degree to capture the contrasts between low and high intensity (arousal 2.81 - 7.10, $M = 4.44$), than the differences in emotional polarity (valence 3.55 - 6.62, $M = 5.37$). In line with studies showing that sensorimotor elements remain more significant for the representation of concrete actions [32], and that the strength of a sensory experience is positively correlated with arousal [19]. These mouth and hand verbs are in the TASA dendrogram divided up between clusters 1 and 3, constituting small and more intense movements as described above, characterized by low to medium arousal (2.60 - 5.60, $M = 3.98$) and enhanced dominance (4.58 - 6.74, $M = 6.12$) respectively.

Verbs describing emotions and facial expressions are grouped together in cluster 2 of the TASA dendrogram, characterized by contrasts between high and low valence (valence 1.96 - 7.89, $M = 3.95$, arousal 2.63 - 6.62 $M = 4.56$). In general abstract concepts have been shown to increasingly rely on affective associations the more abstract they are perceived as being [32]. Emotions and facial expressions are in the HAWIK dendrogram divided into three groups, where cluster 2 captures contrasts in emotional polarity ranging from the negative verbs ‘leer’, ‘sneer’, ‘mock’ and ‘stare’, to the positive verbs ‘smirk’ and ‘laugh’ (valence 3.30 - 7.56, $M = 4.45$; arousal 4.36 - 6.62 $M = 4.7$), while cluster 3 is grouping facial expressions of low intensity in the verbs ‘frown’, ‘glare’, ‘yawn’, ‘glance’, ‘gaze’ and ‘smile’ (arousal 2.63 - 4.52, $M = 3.83$; valence 3.35 - 7.89, $M = 5.7$). Thus grouping verbs characterized by both high valence and low arousal values, which when combined have been found to strengthen motivational approach behavior [33]. Whereas the HAWIK cluster 4 isolates the most negative verbs ‘hate’, ‘sob’, ‘fear’, ‘dread’ and ‘groan’ defined by values at the low end of the valence dimension (valence 1.96 - 3.90, $M = 2.93$; arousal 4.17 - 6.26, $M = 4.89$). Previous studies have established that the three dimensions are not orthogonal, as high or low levels of both valence and dominance are rated as more arousing the further away from neutral they are perceived as being [19]. That is, the hierarchical clustering of emotional verbs thus seem to separate unpleasant feelings rated low in valence as in the verbs ‘sob’, ‘fear’, ‘dread’ and ‘groan’, from relaxed expressions rated low in arousal like ‘smile’ ‘frown’, ‘glance’ and ‘gaze’.

Phonetic parameters

To analyze whether latent semantic parameters of size and intensity might be reflected in the phonetic building blocks of action verbs, their primary stress vowels were spatially mapped out in an articulatory space according to the international phonetic alphabet IPA, framed by their corresponding acoustical F1 and F2 formant frequencies [34]. Although speech sounds constitute a multidimensional feature space, we primarily perceive phonemes by the way in which they are articulated and secondarily by where they are produced in the vocal tract. Meaning, that the main contrast is between sonorant vowels versus plosive and fricative obstruents [35]. In the present study, the hierarchically clustered action verbs describing small size object manipulation appear characterized by high frontal vowels as in ‘P IH1 K’, ‘P IH1 N CH’, ‘S K W IY1 Z’, ‘W IH1 NGK’ and ‘L IH1 K’, while more forceful actions required for moving or transforming objects incorporate phonemes produced further back as in ‘P UH1 L’, ‘P UH1 SH’, ‘P OW1 K’, ‘CH UW1’, ‘CH AA1 P’, ‘S W AA1 L OW’ and ‘TH R OW1’ (Fig. 7). That is, the small size gestures are acoustically defined by high F2 formant frequencies that are maximally dispersed from the F1 formants. While more forceful actions are articulated by back vowels and diphthongs which acoustically have a smaller gap between the F2 and F1 formant frequencies.

In articulatory terms, as the tongue is gradually raised towards the palate, the sonorants undergo a phase shift when the airflow turns turbulent and vowels are transformed into fricatives and plosive stops. Among the action verbs in this study, small movements such as ‘P IH1 K’ and ‘P EH1 K’ are

characterized by both plosive attacks as well as stops which abruptly cut off the resonance of the sonorants. More forceful movements incorporate affricates which begin as plosives and release as fricatives like in ‘CH UW1’ or ‘CH AA1 P’. Whereas verbs initiated by fricatives create a feeling of sustained tension by forcing the flow of air over the edge of the teeth as in ‘S K R EY1 P’ and ‘S AH1 K’. Likewise aspects of contact as in ‘S K W IY1 Z’ and ‘P UH1 SH’ are emphasized by fricatives sustaining the airflow towards the end of the verb. It has been suggested that the phonetic building blocks of plosives, sonorants, and fricatives may have evolved by mimicking the sounds that occur when physical objects collide, resonate or slide across a surface. Interpreted in that sense phonemes may encapsulate additional perceptual characteristics, as verbs beginning with an unvoiced ‘P’ as in ‘P OW1 K’ temporally extends the gap before the diphthong ‘OW’, thereby creating a resonance that acoustically resembles the impact of soft objects with a flexible texture. Whereas the gap before the sonorant is shortened when applying a voiced plosive ‘B’ as in ‘B AY1 T’, resulting in a resonance that acoustically would be associated with collisions of larger more rigid structures [36]. The parameters defined by the dispersion between the F2 and F1 formant frequencies may additionally provide emotional cues that differentiate the hierarchically clustered facial expressions and emotional verbs (Fig. 8). Dynamically lowering the pitch in vowels is perceived as threatening in human speech sounds, while upwards moving formant transitions are to a larger degree associated with positive emotions [11]. Such up- or downward shifts in pitch of the F2 formants are evident in the diphthong double vowels of ‘S M AY1 L’ and ‘G EY1 Z’ moving towards the front ‘IH’ or in ‘F R AW1 N’ towards the back ‘UH’. These contrasts are amplified by transforming the white noise-like fricatives ‘S’, ‘F’ or ‘H’ into alternating closed and open jaw produced sequences of phonemes as in ‘S AA1 B’, ‘F R AW1 N’ and ‘H EY1 T’. Earlier studies have based on latent semantics established that phonaesthemes such as the prefix gl- frequently occurring in verbs like ‘G L AE1 N S’, are not arbitrary phonemes but capture conceptual patterns related to vision [37], that can be traced back to common linguistic Proto-Indo-European roots [38]. Such visual semantic associations are phonetically defined by consonants rather than vowels, but the upwards or downwards formant shifts in ‘G EY1 Z’ or ‘G L EH1 R’, suggests that additional emotional content could be encoded in the phoneme transitions. Likewise, downward frequency shifts due to the lowered F3 third formant characteristic of the liquid consonant ‘R’, are in this study prevalent in negative emotions such as ‘L IH1 R’, ‘F IH1 R’, ‘S N IH1 R’, ‘G L EH1 R’, ‘S T EH1 R’ and ‘S K EH1 R’, that are all rated below 4.50 in perceived valence.

Whether language is seen as grounded in simulation literally dependent on sensorimotor circuits, or rooted in symbolic associations constituted by statistical word representations, there is an emerging consensus on the need to adapt a pluralist view about embodiment and semantics [39] [40]. Spatial parameters have in previous studies been extracted from sentences describing horizontal or vertical movements based on latent semantics retrieved from word co-occurrences [8]. Potentially based on sensorimotor links between perception of shapes and motion, which may have constrained how aspects of size and intensity are mapped onto the consonants and vowels of words [41]. Underlying structural dimensions of size seem hardwired into speech articulation, as grasping objects of increasing size has been shown to simultaneously enlarge both the lip kinematics and mouth aperture when pronouncing vowels [42]. The phonetic contrasts differentiating the verbs in the present study are in line with recent findings related to how phonetic features are represented in the brain when listening to speech, showing that the relation between the F2 and F1 formants constitutes the first principal component in neuronal encoding of vowels. Indicating that a higher order encoding of acoustic formant parameters provide the foundation for neuronal response properties that are maximally tuned to differentiate open low back from close high front vowels [35]. In the present study exemplified by the contrasts between the primary stress vowels in ‘CH AA1 P’ versus ‘P IH1 K’. Likewise when pronouncing consonant vowel syllables, the hierarchical structure of consonants primarily reflects whether they are produced by the lips or dorsal tongue as in ‘P’ or ‘K’ versus coronal tongue as in ‘CH’ or ‘SH’. While the hierarchical structure of vowels at the highest level differentiate unrounded high front and low back vowels as in ‘P IH1 K’ and ‘CH AA1 P’ from rounded high back vowels as in ‘P UH1 SH’ [43]. These articulatory structures appear in the present study reflected in

the latent semantic contrasts between small size motion as in ‘P EH1 K’ , combining labial and dorsal consonants with unrounded vowel, versus forceful movements such as ‘TH R OW1’ combining coronal consonant with rounded vowel. A simplified representation of phonetic features could be interpreted as a continuum, vertically going from open jaw low back vowels such as ‘AA’ ‘ in ‘S W AA1 L OW’ to near close frontal vowels such as ‘IY’ in ‘S K W IY1 Z’, that are in turn horizontally contrasted against high back rounded vowels like ‘UH’ in ‘P UH1 L’. Suggesting, that these articulatory and acoustic dimensions not only maximize phonetic contrasts to facilitate comprehension, but also reflect spatial parameters of intensity and emotional content encoded in action verbs, that can be extracted from word co-occurrences in the surface structure of language.

Methods

Initially 3×20 hand, face and emotion related action verbs were selected, constituting half of the action verbs similarly used in a fMRI functional magnetic resonance neuroimaging study, demonstrating that the selected action verbs activated premotor cortices in the brain during a passive reading task [14]. Latent semantic analysis LSA [44] [15] was applied in order to retrieve an adjacency matrix based on the HAWIK text corpus consisting of 22829 words found in 67380 excerpts of Harvard Classics literature, Wikipedia articles and Reuters news [16]. The cosine similarities of verbs based on the HAWIK text corpora were subsequently compared against those retrieved based on the TASA corpus consisting of 92409 words in 37651 fiction and non-fiction texts extracted from novels, news articles, and other general knowledge reading material that the average American student has been exposed to from 3rd grade until reaching first year of college [17]. Using singular value decomposition SVD to reduce dimensionality [45], the original $m \times n$ term-document matrix \mathbf{X} is decomposed into a product of three other matrices:

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$$

where the \mathbf{U} matrix, similar to the original matrix has m rows of words, while the columns now consist of r eigenvectors representing the principal components in the data. Likewise the transpose of the orthonormal matrix \mathbf{V}^T has as before n columns of documents but now related to r rows of eigenvectors or principal components. The very purpose of the decomposition is to scale down the number of parameters based on a $\mathbf{\Lambda}$ square matrix containing r singular values λ arranged along the diagonal in decreasing order, which as eigenvalues scale the eigenvectors of the rectangular matrices to each other and thereby derive a matrix of reduced dimensionality:

$$\mathbf{Z}_k = \mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}_k^T$$

where only the k largest singular values of the $\mathbf{\Lambda}$ diagonal matrix are retained. As a result the number of parameters in the rectangular \mathbf{U}_k and \mathbf{V}_k^T matrices are reduced to what would correspond to the principal components containing the highest amount of variance in the matrix. Thus allowing us to reconstruct the original input based on a \mathbf{Z}_k matrix of lower dimensionality which is embedding the underlying structure of the data. Geometrically speaking, the terms and documents in the condensed \mathbf{Z}_k matrix can be interpreted as points in a k dimensional subspace, which enables us to calculate the degree of similarity between matrices based on the dot or inner product of their corresponding vectors. Interpreting the matrix multiplication geometrically the cosine similarity between two words represented by their vectors can be expressed as

$$\cos \theta = \frac{x \cdot y}{\|x\|\|y\|}$$

where $x \cdot y$ signifies the dot product of the vectors, and $\|x\|\|y\|$ the Euclidean norm corresponding to the

square root of the dot product of each vector with itself.

The optimal number of 300 dimensions used for the LSA analysis based on the TASA has been determined based on a synonymy test [15]. To determine the optimal number of dimensions for the HAWIK corpus a similar synonymy test was implemented, which based on questions from the TOEFL ‘test of english as a foreign language’ compared the LSA cosine similarity of the multiple choice test synonyms, while varying the number of eigenvectors until an optimal percentage of correct answers were returned. For the HAWIK matrix the optimal result of 71,2% correctly identified synonyms based on LSA was found when reducing the singular value decomposition SVD to the most significant 125 eigenvalues. This result is above the 64.5% TOEFL average test score achieved by non-native speaking US college applicants, on par with previous results obtained using either LSA or probabilistic topic models [46].

To identify significant structures among the action verbs, hierarchical clustering was applied to the two adjacency matrices derived from the HAWIK and TASA corpora, using Pearson correlation as distance measure. Assessing to what degree the resulting hierarchical tree structures are supported by the data, approximately unbiased AU p-values were computed by multistep multiscale resampling changing the sample size of the bootstrap replicates [18]. Selecting the action verbs which were hierarchically clustered similarly based on both the HAWIK and TASA adjacency matrices, the words were annotated with their corresponding user rated word norms available online from the “Norms of valence, arousal and dominance for 13915 English lemmas” data set [19]. Similar to the smaller ANEW data set [29] the user ratings assess how pleasant, intense and controlled the verbs are described as being on a scale from 1 to 9, along the psychological dimensions of valence, arousal and dominance [20].

Subsequently the action verbs were transformed into ARPAbet phonemes using the CMU text to speech pronunciation dictionary [21], and their primary stress vowels projected into an articulatory space defined by tongue height and front-back position in the international phonetic alphabet (IPA), to identify common spatial parameters acoustically framed by the corresponding average F1 and F2 formant frequencies [22].

References

1. Deacon TW (1998) The symbolic species: the co-evolution of language and the brain. W. W. Norton & Company.
2. Glenberg AM, Gallese V (2011) Action-based language: A theory of language acquisition, comprehension, and production. *Cortex* doi:10.1016/j.cortex.2011.04.01: 1-18.
3. Lakoff G, Johnson M (1999) Philosophy in the flesh; the embodied mind and its challenge to western thought. Basic Books.
4. Engel AK, Maye A, Kurthen M, König P (2013) Where’s the action ? the pragmatic turn in cognitive science. *Trends in Cognitive Sciences* 17.
5. Meier BP, Robinson MD (2004) Why the sunny side is up - associations between affect and vertical position. *Psychological Science* 15.
6. Barsalou LW (2008) Grounded cognition. *Annual Review of Psychology* 59: 617-645.
7. Barsalou LW, Santos A, Simmons WK, Wilson CD (2008) Language and simulation in conceptual processing. *Symbols, embodiment and meaning* : 245-283.
8. Louwerse MM (2011) Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science* 3: 273-302.

9. Louwerse MM, Zwaan RA (2009) Language encodes geographical information. *Cognitive Science* 33: 51-73.
10. Louwerse MM, Benesh N (2012) Representing spatial structure through maps and language: Lord of the rings encodes the spatial structure of middle earth. *Cognitive Science* 36: 1556-1569.
11. Myers-Schulz B, Pujara M, Wolf RC, Koenigs M (2013) Inherent emotional quality of human speech sounds. *Cognition and emotion* 27: 1105-1113.
12. Maurer D, Pathman T, Mondloch CJ (2006) The shape of boubas: sound-shape correspondences in toddlers and adults. *Developmental Science* 3: 316-322.
13. Monaghan P, Christiansen MH, Chater N (2007) The phonological-distributional coherence hypothesis: cross-linguistic evidence in language acquisition. *Cognitive Psychology* 55: 259-305.
14. Moseley R, Carota F, Hauk O, Mohr B, Pulvermüller F (2011) A role for the motor system in binding abstract emotional meaning. *Cerebral Cortex* doi:10.1093/cercor/bhr238: 1-14.
15. Landauer TK, Dumais ST (1997) A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104: 211-240.
16. Petersen MK (2012) LSA software & HAWIK corpus matrices. Technical University of Denmark <https://dl.dropboxusercontent.com/u/5442905/LSA.zip>.
17. Landauer TK, McNamara DS, Dennis S, Kintsch W (2007) Handbook of latent semantic analysis. University of Colorado <http://lsa.colorado.edu>.
18. Shimodaira H (2004) Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *The Annals of Statistics* 32.
19. Warriner AB, Kuperman V, Brysbaert M (2013) Norms of valence, arousal and dominance for 13915 english lemmas. *Behavior Research Methods* : 1-17.
20. Russell JA (1980) A circumplex model of affect. *Journal of Personality and Social Psychology* 39: 1161-1178.
21. CMU (1976) The cmu pronouncing dictionary. Technical report, Carnegie Mellon University.
22. Catford JC (1988) A practical introduction to phonetics. Clarendon Press.
23. Graziano MS, Taylor CS, Moore T, Cooke DF (2002) The cortical control of movement revisited. *Neuron* 36: 349-362.
24. Jeannerod M, Arbib MA, Rizzolati G, Sakata H (1995) Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neurosciences* 18.
25. Konkle T, Oliva A (2012) A real-world size organization of object responses in occipitotemporal cortex. *Neuron* 74: 1114-1124.
26. Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Neuroscience* 10: 186-198.
27. Moreno-Dominguez D, Anwender A, Knösche TR (2014) A hierarchical method for whole-brain connectivity-based parcellation. *Human Brain Mapping* DOI: 10.1002/hbm.22528.
28. Murphy GL, Andrew JM (1993) The conceptual basis of antonymy and synonymy in adjectives. *Journal of memory and language* 32: 301-319.

29. Bradley MM, Lang PJ (2010) Affective norms for english words (anew): Stimuli, instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida.
30. Lang PJ, Bradley MM (2010) Emotion and the motivational brain. *Biological Psychology* 84: 437-450.
31. LeDoux J (2012) Rethinking the emotional brain. *Neuron* 73: 653-676.
32. Kousta ST, Vigliocco G, Vinson DP, Andrews M, Campo ED (2011) The representation of action words: why emotion matters. *Journal of Experimental Psychology* 140: 14-34.
33. Recio G, Conrad M, Hansen LB, Jacobs AM (2014) On pleasure and thrill: the interplay between arousal and valence during visual word recognition. *Brain & Language* 134: 34-43.
34. Ladefoged P (1989) Representing phonetic structure. *Working Papers in Phonetics* 73.
35. Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. *Science* 343.
36. Changizi MA (2011) *Harnessed - how language and music mimicked nature and transformed ape to man*. BenBella Books.
37. Otis K, Sagi E (2008) Phonaesthemes: a corpus based analysis. In: 30th Annual Conference of the Cognitive Science Society. pp. 65-70.
38. Boussidan A, Sagi E, Ploux S (2009) Phonaesthetic and etymological effects on the distribution of senses in statistical models of semantics. In: *Distributional semantics beyond concrete concepts*. CogSci, pp. 35-40.
39. Meteyard L, Cuadrado SR, Bahrami B, Vigliocco G (2012) Coming of age: a review embodiment and the neuroscience of semantics. *Cortex* 48: 788-804.
40. Pulvermüller F (2013) How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences* 17: 458-470.
41. Ramachandran V, Hubbard E (2001) Synaesthesia - a window into perception thought and language. *Journal of Consciousness Studies* 8: 3-34.
42. Gentilucci M, Corballis MC (2006) From manual gesture to speech: a gradual transition. *Neuroscience and Biobehavioral Reviews* 30: 949-960.
43. Bouchard KE, Mesgarani N, Johnson K, Chang EF (2013) Functional organization of human sensorimotor cortex for speech articulation. *Nature* 495: 327-331.
44. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman RA (1990) Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41: 39-407.
45. Furnas GW, Deerwester S, Dumais ST, Landauer TK, Harshman RA, et al. (1988) Information retrieval using a singular value decomposition model of latent semantic structure. In: 11th Annual International ACM SIGIR Conference. pp. 465-480.
46. Griffiths TL, Steyvers M, Tenenbaum JB (2007) Topics in semantic representation. *Psychological Review* 114: 211-244.